Brevity is the soul of Twitter: The constraint affordance and political discussion

Kokil Jaidka
Wee Kim Wee School of Communication and Information
Nanyang Technological University

Alvin Zhou
Annenberg School for Communication
University of Pennsylvania

Yphtach Lelkes*
Annenberg School for Communication
University of Pennsylvania

Author Note

Corresponding Author: Yphtach Lelkes, ylelkes@upenn.edu. Paper forthcoming in the Journal of Communication

Abstract

Many hoped that social networking sites would allow for the open exchange of information and a revival of the public sphere. Unfortunately, conversations on social media are often toxic and not conducive to healthy political discussion. Twitter, the most widely used social network for political discussions, doubled the limit of characters in a Tweet in November 2017, which provided a natural experiment to study the causal effect of technological affordances on political discussions with a discontinuous time series design. Using supervised and unsupervised natural language processing methods, we analyze 358,242 Tweet replies to U.S. politicians from January 2017 to March 2018. We show that doubling the permissible length of a Tweet led to less uncivil, more polite and more constructive discussions online. However, the declining trend in the empathy and respectfulness of these tweets raises concerns about the implications of the changing norms for the quality of political deliberation.[1]

Keywords: Political Communication, Political Discussion, Social Media, Computational Social Science, Affordances

Brevity is the soul of Twitter: The constraint affordance and political discussion

Over the past decade, we have witnessed a shift in political communication. In the context of political campaigning, social media platforms such as Twitter and Facebook now provide politicians with a platform to mobilize and engage in meaningful discussion with their constituents (Theocharis, Barberá, Fazekas, Popa, & Parnet, 2016). At the same time, these platforms give citizens the ability to access political information, participate in local events and citizen movements, and voice their support or dissent against their government. Social networking sites are now essential tools in the democratic and civic process, as they offer a neutral and open platform for dialogue, act as the bedrock where citizens build ideas about politics through everyday talk (Wojcieszak & Mutz, 2009; Wyatt, Katz, & Kim, 2000), and facilitate direct contact between citizens and their political representatives (e.g., Liu & Zhang, 2013; Tufekci & Wilson, 2012).

Everyday political discussions on social networking sites are considered to be one of the many pathways to political participation (for a detailed review, see Cook, Carpini, & Jacobs, 2007; Eveland, Morey, & Hutchens, 2011; Hopmann, Matthes, & Nir, 2015). However, the question of whether 'meaningful' political discussions are possible on social networking sites, depends on whether or not it is possible to have 'rational' discussions through linguistic exchanges (Habermas, 1984), or being able to communicate ideas and build consensus in a polite and respectful manner (e.g., Liu & Zhang, 2013; Mutz & Reeves, 2005; Papacharissi, 2004). Some researchers are skeptical of the potential of social networking sites as public spheres for political deliberation (Bail et al., 2018; Mendelberg, 2002; Schkade, Sunstein, & Kahneman, 2000). One of the concerns raised is that political discussions on platforms such as Twitter are of low quality since they are often toxic, provocative, and replete with trolls who incite hyperbole (e.g., Berry & Sobieraj, 2013; Nithyanand, Schaffner, & Gill, 2017; Theocharis et al., 2016).

Many explanations for the low quality of online discourse have been put forth, including the existence of disinformation campaigns, the polarization of politics, spurious

accounts, and echo chambers (e.g., Berry & Sobieraj, 2013; Chen, 2017). Incivility ultimately limits users' ability to have meaningful and constructive exchanges with each other or with their political representatives (Theocharis et al., 2016).

We argue that another factor that regulates the quality of political discussions are the technological affordances for communication. Affordances are properties that enable or constrain the potential for action (Faraj & Azad, 2012). Examples of affordances in digital communication include the ability to have real-time discussions, to hide one's true identity and to navigate information in a hierarchical manner (Friess & Eilders, 2015; Sundar, 2008). A number of studies have shown that the association of one affordance (anonymity) with the quality of online political discussions (e.g., Halpern & Gibbs, 2013; Theocharis, Lowe, van Deth, & García-Albacete, 2015; Towne & Herbsleb, 2012).

In this paper, we examine whether another affordance—constraints on message length—impacts the quality of online political discussion. Using an interrupted times series design combined with an automated content analysis of over three hundred and fifty thousand tweets involving politics, we compare various message features associated with ideal political discussion before and after Twitter moved from allowing 140 characters per tweet to 280 characters. We show that Twitter's relaxation of its message length constraint affected some, but not all messages features which improved the quality of political discussion. For instance, while the changeover increased the prevalence of less uncivil, more formal and more constructive messages, there was a decrease in the empathy and respectfulness of messages. We also test the robustness of our results using different model specifications, bandwidths, and a placebo test.

Our theoretical contribution is twofold. First, we extend prior research by shifting attention to social media affordances that ultimately influence the quality of discussions on social media. Changes to the design of platforms, which are more tractable than, for instance, censoring content, have substantive effects on the health of online political discussion. Second, by revealing the trade-offs in content and style that users face when

articulating their opinion, our study yields valuable insights about how the use of these platforms may affect the quality of public discourse and users' experiences of online political participation as a more reactive versus a more deliberative exercise. We also offer a methodological contribution in the form of language models that can automatically label text according to its uncivil and deliberative qualities.

The Twittersphere

Political discussions on Twitter and other online platforms allow people with diverse perspectives and opinions to participate in casual conversation. In theory, they can cut across diverse social networks In addition to facilitating horizontal discussion between citizens, platforms like Twitter allow vertical discussions between citizens and policymakers (e.g., Ausserhofer & Maireder, 2013; Davis, 2010; Effing, van Hillegersberg, & Huibers, 2011).

Twitter's users organically create a networked sphere of political discussion which is structurally independent of the traditional arena of politics or news; yet, it connects with the two through official affiliations and real-life interactions (Lindgren & Lundström, 2011). The online political sphere has the potential to reinvigorate offline politics by allowing millions of individual contributions, subverting the often monolithic agenda set by traditional mass media and policymakers in the offline world (e.g., Habermas, Lennox, & Lennox, 1974; Papacharissi, 2010; Shirky, 2008).

The fact that Twitter facilitates discussion between citizens and policymakers does not, in itself, make it a "democratic utopia" (Papacharissi, 2004; Stroud, Scacco, Muddiman, & Curry, 2015). Online political discussions rarely meet the ideals for political deliberation, such as open communication, equality, inclusivity and compromise (e.g., Friess & Eilders, 2015; Nithyanand et al., 2017; Theocharis et al., 2016; Wojcieszak, 2010). Political discussions on Twitter are often replete with toxic and abusive responses, flaming, and group-based stereotyping (Halpern & Gibbs, 2013; Theocharis et al., 2016). Users participating in discussions on Twitter are found to be unlikely to indulge in reflection, or

frame coherent arguments, which negatively impacts the quality of political discussions (Janssen & Kies, 2005; Stromer-Galley & Martinson, 2009). A study by Theocharis et al. (2016) argued that users' toxic and uncivil responses were responsible for shutting down any meaningful political engagement between elected U.S. politicians and the citizenry.

To summarize, many studies have identified a gap in the normative ideals of Twitter as an ideal platform for lively and inclusive political debates and its role in actually facilitating political deliberation. The following section contextualizes the study's hypotheses in the current understanding of how Twitter's technological features, or affordances, enable deliberation and politeness or constrain incivility.

Technological Affordances and the Potential for Civic Discussion

When designing online platforms, developers make many choices that change "possibilities for action between an object/technology and the user that enables or constrains potential behavioral outcomes in a particular context" (Evans, Pearce, Vitak, & Treem, 2017, p. 36). These possibilities for action are often called affordances (Gibson, 1977; Greeno, 1994) and can be broadened or narrowed by changing the underlying technical specifications. For instance, in the context of political discussions, Twitter affords the scope for two-way communication between politicians and citizens because of how asymmetrical friendship relationships are specified by default (Grant, Moon, & Busby Grant, 2010). The affordance for two-way communication is restricted on Facebook, where, in the default setting, two individual users may need to "follow" each other in order to send a message or even view each other's profile.

Friess and Eilders (2015) and Janssen and Kies (2005) identify some affordances which affect political deliberation. Anonymity, for instance, affords higher participation but uncivil discourse (Towne & Herbsleb, 2012). Real-time participation in synchronous chats provokes instantaneous reactions but reduces reflexive, coherent argumentation and rebuttal (Janssen & Kies, 2005; Stromer-Galley & Martinson, 2009). Flattened follower-followee connections overcome the depersonalizing effects of digital communication

and increase citizens' emotional closeness to political elites and elected candidates (Lee & Oh, 2012).

Navigable information interfaces support political deliberation by fostering clear communication, rational argumentation, and constructiveness (Towne & Herbsleb, 2012). Among the studies exploring the role of affordances in political deliberation, a majority have focused on how the anonymity and deindividuation reduces the rationality, sincerity, and civility of the conversation (e.g., Halpern & Gibbs, 2013; Theocharis et al., 2015; Towne & Herbsleb, 2012); on the other hand, anonymity also improves the likelihood of participation and hence the quantity and inclusivity of political debates (Towne & Herbsleb, 2012).

At least two studies have examined the relationship between message length (another affordance, and our particular focus) and political discussion quality. However, these studies turn up conflicting findings: Papacharissi (2004) reported that longer messages posted in political discussions were significantly more uncivil than shorter messages. Other publications have argued that Twitter's character limit constraints can encourage individual creativity and improve content quality (Gligorić, Anderson, & West, 2018; Joyce, 2009). On the other hand, Oz, Zheng, and Chen (2018) found that shorter messages on Twitter were more uncivil, impolite and less deliberative than longer messages on Facebook; however, in their experimental replication, they observed only a significant improvement in deliberation in the longer messages.

These studies offer important insights; however, they are limited in crucial ways. For instance, Papacharissi (2004) does not account for self-selection and the possibility that those that tend to write longer messages may already behave more uncivilly. While the Oz et al. (2018) experiment solves this internal validity issue, the character limit is confounded with other characteristics and norms of the platform. That is, subjects responded to either a mock Facebook post or a mock Twitter post. Given these conflicting findings, we begin with a research question:

RQ. What effect does doubling the character limit have on incivility?

Character limits may also limit the formality of discourse, making it less amenable to clear and open communication. As the original Twitter users embraced its character limit, they organically devised a series of conventions to convey meaning and information structure: slang, netspeak, and abbreviations were adopted to express reactions (LOL, SMH), share opinion (IMHO, AFAIK), reference other users (-mentions, h/t, via), label topics (hashtags), and identify propagated messages (RT, QT). However, while the need for concision required adherence to convention, this may lead to the sacrifice of clarity over the content. Twitter users were constrained to expressing themselves in 140 characters, so they were more likely to use slang and abbreviations to compress their ideas, as compared to their behavior on platforms such as Facebook which do not have such constraints (Jaidka, Guntuku, Buffone, Schwartz, & Ungar, 2018; Lin & Qiu, 2013). With the availability of more characters, however, users are more likely to apply a less casual and more professional writing style, as has been reported in some previous studies (Gligorić et al., 2018; Jaidka et al., 2018). Accordingly, we posit the following hypothesis:

H1: Comments posted on Twitter after the 280-character limit change will be less informal (more polite) than before the change.

Civility is not a sufficient condition for achieving the "democratic potential" of the internet. Political discussions also need to be substantive, in terms of providing evidence for their claims and raising arguments that aim to build consensus or move the conversation forward. However, Papacharissi (2004) raises the concern that greater adherence to politeness could restrict political debate by substituting it with a formal exchange of rhetoric. With fewer stakes in the argument, participants may be less likely to take a stance, provide pieces of evidence or persuade others. Instead of simply being civil, political discussion should aim to encompass "conversation specifically aimed at political action," and make comprehensive arguments validated with facts and data (Freelon, 2010; Halpern & Gibbs, 2013; Papacharissi, 2004; Stroud et al., 2015; Theocharis et al., 2016).

Chen (2017) speculates that both incivility and political deliberation can increase when message length constraints are relaxed; however, it is not known whether this effect would remain in the absence of the deindividuation differences between platforms. The study by Oz et al. (2018) reported with observational and experimental analysis that political messages addressed to the U.S. White House on Twitter were significantly less likely to mention factual data as compared to Facebook.

Besides high-quality argumentation, political "discussions" also imply an *exchange* of arguments; consequently, participants are required not simply to speak but also to listen and respond (Barber, 1984). Friess and Eilders (2015), for instance, emphasize the importance of interactivity as a key dimension of deliberative behavior. Other studies examining interactivity in deliberation, such as Habermas (1990), Walther, Gay, and Hancock (2005), Stroud et al. (2015), and Himmelroos (2017) have stressed the importance of role-taking and empathy in a rational discussion.

To summarize, in general we anticipate that with more characters at their disposal, users can afford to be more civil, construct better arguments and be more pleasant to each other in their political discussions. This is based on empirical findings from the many papers that have compared the quality of political deliberation across social networking sites with different affordances (Chen, 2017; Oz et al., 2018; Stroud et al., 2015), as well as previous studies that have theorized that improving the opportunities to participate (e.g., designing social media platforms with better affordances) would lead to an increase or improvement in deliberative behavior (Cook et al., 2007; Neblo, Esterling, Kennedy, Lazer, & Sokhey, 2010). We anticipate that extending the character limit may afford users more space to make a cogent argument or support their views with evidence, thus facilitating more sophisticated and deliberative political discussions (Chen, 2017; Theocharis et al., 2016). Accordingly, we posit our second hypothesis in two parts:

H2a: Comments posted on Twitter after the 280-character limit change will comprise a better quality of arguments than before the change.

H2b:Comments posted on Twitter after the 280-character limit change will comprise a better quality of interactions than before the change.

Data and Methods

We built our dataset of political discussions on Twitter by adopting the method of Theocharis et al. (2016) to identify the Twitter replies to 536 U.S. Congressmen and Congresswomen who were in office before November 7, 2017, when Twitter instituted a change in the character limit. Using the Twitter replies to members of Congress increases the odds that a particular tweet is part of a political discussion. We filtered a Twitter 1% sample, collected using the Twitter streaming API between January 2017 - March 2018, to retain all replies (i.e., tweets starting with "politicians") to these set of Twitter handles, occurring between January 1, 2017, and March 31, 2018. These dates give us a larger pre-intervention and post-intervention observation period and the ability to assess whether our results remain stable if we consider different bandwidths in the regression equation.

In Table 1, we provide a summary of our dataset with the number of tweets remaining after each pre-processing step. First, we performed language filtering to retain only English-language tweets. Next, we removed any tweets which were retweets and thus did not constitute an actual reply. We manually inspected the character limits of the tweets and found that because of the differences in encoding, the actual length of tweets in the pre-intervention period was often up to 145 characters. Accordingly, we considered 145 characters as a better approximation to compliance than 140 characters for future experiments. We removed tweets by the 1% users of the population who were subjected to the intervention early (the 'early-access' users) on September 27. Finally, our dataset comprises 358,242 replies respectively to U.S. politicians from January 2017 - March 2018.

Table 1 also provides statistics on the unique number of users in our dataset. We also identify the *compliers*: users who tweeted more than 145 characters at least once after the character limit change, and the *non-compliers*: users who, despite the extension, never tweeted more than 145 characters. Wilcoxon signed-rank test shows that compliers and

non-compliers differ significantly across some of their account characteristics, such as the number of followers. Non-compliers were more likely to have been Twitter users for longer ($\mu = 4.43$ years) but compliers had a higher average number of tweets posted ($\mu = 31681.93$, SD = 52822.50) and a higher per-user occurrence in our dataset ($\mu = 2.82$, SD=3.78) as compared to non-compliers ($\mu = 29408.74$, SD = 52030.67; $\mu = 2.11$, SD=2.13), which suggests that compliers might be more active on Twitter and already more engaged in politics than non-compliers. A comprehensive comparison of compliers and non-compliers is provided in the Supplementary Materials. Our within-subjects model specification controls for these individual differences, and our results remain substantively similar to our other methods.

Operationalization. Based on previous work in analyzing political discourse, civility, and deliberation on social media, we have identified a set of language features that can capture the quality of political discussions. Specifically, we focus on whether the tweets reflect uncivil or deliberative behavior. Furthermore, we distinguish the criteria followed to measure the quality of its arguments, i.e., the use of justification and constructiveness, from the criteria to measure the qualities of interactivity as per Friess and Eilders (2015), i.e., reciprocity, and empathy and respect. These are elaborated below. We have also described how they fit in with our hypotheses and research question, in Table 2:

- We operationalize incivility in terms of the daily number of uncivil tweets, or the average percentage proportion of uncivil words per tweet as predicted by different machine learning and dictionary-based lexical methods.

- We operationalize formality in terms of the decrease in the use of informal words or increase in politeness in political discussions.

- We operationalize a measurement of deliberative quality in terms of the 'justification' and the 'constructiveness' attributes of political discussions, based on previous work which has used three criteria: the presence of supporting or external materials, the

use of 'comprehensive' arguments and the use of numbers or statistics (Oz et al., 2018; Papacharissi, 2004; Stroud et al., 2015). Other work by Esteve Del Valle, Sijtsma, and Stegeman (2018) and Rowe (2015) have also identified the use of personal anecdotes, values, and ideologies to support claims. Friess and Eilders (2015) and Steenbergen, Bächtiger, Spörndli, and Steiner (2003) consider arguments to be constructive if, besides providing factual arguments, they also attempt to resolve conflicts or propose solutions.

  - **Justification:** Reflects whether the author offers a justification, either based on personal experiences, values, and feelings or data, links, and facts (Esteve Del Valle et al., 2018; Oz et al., 2018; Rowe, 2015).
  - **Constructiveness:** Reflects the author's attempt to move the conversation forward, build and bring about consensus, and resolve conflicts by pointing out facts, identifying common ground or proposing solutions (Friess & Eilders, 2015; Steenbergen et al., 2003).

- Finally, we operationalize interactivity in two ways: Reciprocity, and Empathy and Respect. Reciprocity is defined in terms of writing that asks a genuine question and invites responses in the form of answers or further information (Stromer-Galley, 2007; Stroud et al., 2015). We operationalize empathy and respect based on prior criteria to measure discursive quality (Esteve Del Valle et al., 2018; Steenbergen et al., 2003) as positive comments that are sensitive or empathetic to others' viewpoints, and respectful towards other discussants.

  - **Reciprocity:** Reflects whether the author engages with others in a substantial rather than a formal manner, with a tweet intended to elicit a response or further information (Friess & Eilders, 2015; Rowe, 2015; Stroud et al., 2015).
  - **Empathy and Respect:** Reflects the author's acknowledgment of or sensitivity to others, manifested in positive comments, an empathetic or a

respectful response acknowledging other viewpoints (Esteve Del Valle et al., 2018; Steenbergen et al., 2003).

As a first step, we trained a supervised machine learning classifier on 6000 hand-annotated tweets, so that it could label each tweet in our dataset according to whether it was relevant to politics. Labeling tweets as either being relevant to politics (1) or not (0) helped us to ensure that our inferences about political discussions were valid.

Uncivil behavior was measured in terms of the use of name-calling, profanity, hate speech or invocation of stereotypes of a homophobic, racist, sexist or xenophobic nature (Chen, 2017). Communication scholars have argued that incivility is more than offensiveness. For example, Papacharissi (2004) defined incivility as impolite behavior that threatens democracy and has lasting repercussions on the common good. In other words, uncivil tweets must be offensive, but offensive tweets are not necessarily uncivil. Papacharissi's (2004) conceptualization was adopted by recent studies using human coding (e.g., Groshek & Cutino, 2016; Muddiman, McGregor, & Stroud, 2018; Rowe, 2015; Santana, 2014). In this study, we have applied and compared four different language-based models of incivility:

1. We trained an incivility classifier on a dataset of 6000 hand-annotated tweets from our dataset labeled as uncivil (1) or civil (0) and captured a wide range of abusive, racist, threatening or exaggerative behavior common to online political incivility Stroud et al. (2015).

2. We trained an offensiveness classifier on a dataset of hand-annotated tweets provided by Davidson, Warmsley, Macy, and Weber (2017) which labels tweets as offensive (1) or inoffensive (0).

3. We applied the uncivil words dictionary hand-annotated from a corpus of New York Times comments by Muddiman et al. (2018).

4. We applied the swear words dictionary from Linguistic Inquiry and Word Count (LIWC) (Pennebaker, Booth, Boyd, & Francis, 2015), which measures the percentage proportion of swear words in a text.

Additionally, some scholars have considered politeness and formality to be essential qualities for constructive political debate (Chen, 2017; Papacharissi, 2004). Politeness refers to adherence to etiquette and an extension of courtesy to fellow discussants (Papacharissi, 2004). We used the Stanford Politeness Application Programming Interface (API) to access a language model provided by Danescu-Niculescu-Mizil, Sudhof, Jurafsky, Leskovec, and Potts (2013) which scores each tweet on a scale of 0 to 1 for its politeness. Informality (or, conversely, the lack of formality) is used to reflect messages with less substance and is considered a less effective means of political discussion. We used the informality dictionary provided by LIWC (Pennebaker, Booth, et al., 2015) which defines informality as content that is relatively higher in its use of assents, fillers, swear words and netspeak.

Following the original definition by Gastil (2008), we sought to measure deliberative behavior in terms of the use of analytical and social language to build consensus, through facts and supporting rationales, raise and answer questions, consider and rationalize alternatives and express positions on issues. To our knowledge, there are no supervised machine learning classifiers available to label tweets according to their deliberative content automatically. Accordingly, we trained classifiers on a set of hand-annotated tweets and used them to label the entire dataset.

Almost all of the operationalizations encode the presence or absence of a feature. Politeness is output as a continuous variable constituting a likelihood percentage. The dictionary-based methods used to measure informality and uncivil words comprise normalized, within-tweet percentages indicating the proportion of words which were representative of the category of interest. For example, a score of 3 for informality would imply that if a tweet comprised a hundred words, then there would be three among those hundred that were informal words.

Validation. For the incivility and the deliberative classifiers developed in this study, machine learning methods were implemented to train logistic regression classifiers on the language of 6000 tweets labeled by four annotators. The classifiers use a weighted function of the normalized frequency distribution of words and phrases in a tweet to assign a 0 or 1 label for the presence of incivility and each deliberative attribute. We validated the classifiers against held-out labeled data in ten-fold cross-validation and obtained an average accuracy of 79% across all the categories. Details on the inter-coder agreement and the predictive performance are provided in Table 4 and 5 of the Supplementary Materials.

For measuring politeness and offensiveness, we used supervised machine learning models trained on hand-annotated data developed by computer scientists (Danescu-Niculescu-Mizil et al., 2013; Davidson et al., 2017). The politeness API has also been extensively used by the computer science research community (e.g., Althoff, Danescu-Niculescu-Mizil, & Jurafsky, 2014; Jongeling, Sarkar, Datta, & Serebrenik, 2017; Tan, Niculae, Danescu-Niculescu-Mizil, & Lee, 2016) in linguistic analyses of social media text. The classifier and the annotated dataset on offensiveness have been validated in subsequent studies (Almeida, Souza, Nakamura, & Nakamura, 2017; Olteanu, Talamadupula, & Varshney, 2017).

For measuring swear words and informality, we used the dictionaries provided by Linguistic Inquiry and Word Count (LIWC) 2015 (Pennebaker, Boyd, Jordan, & Blackburn, 2015), a computerized program developed by psychologists to automatically categorize words in a text (Pennebaker, Boyd, et al., 2015). Dictionaries of LIWC have been validated in subsequent language analyses of social media posts.[2] Further details about the supervised and unsupervised methods used to mine these features are provided in the Supplementary Materials.

Interrupted Time Series (ITS) Regression Model

Our primary approach was an interrupted time series analysis (ITS) to determine whether the character-limit change induced an improvement in political discussions. ITS is

a variation of regression discontinuity designs (RDD) where the running variable is time (Lopez Bernal, Cummins, & Gasparrini, 2017). This approach was ideal because Twitter data is time-stamped, with a high frequency of daily measurements and a well-defined moment of intervention. ITS design requires a clear differentiation of the pre-intervention and post-intervention period (Lopez Bernal et al., 2017)—the extension of character limit from 140 to 280 on November 7, 2017, in our case. The unit of analysis in this set of regressions was the daily mean score of an uncivil or a deliberative attribute. The quantity of interest was the immediate change in the different attributes measuring uncivil and deliberative behavior after November 7, 2017. To tackle the inconsistency of the numbers of tweets on each day and avoid Type I errors, we bootstrapped the analysis for 100 iterations, sampling the approximate daily mean of 700 tweets per day for each analysis and reported the average effect sizes and standard errors across all the iterations. All variables were scaled to values between 0 to 100 and mean-centered before analysis in order to more readily interpret the interaction term. The independent variable was dichotomous, indicating whether it was before or after the intervention. The number of days since the intervention is represented by T in the following ordinary least squares models:

$$Y_{feature,t} = \beta_0 + \beta_1 T + \beta_2 X_t + \beta_3 T X_t \tag{1}$$

In the above model, T is the relative time distance from November 7, 2017. For example, *time* equals to -3 on November 4, 2017 and equals to 7 on November 14, 2017. X is a dummy variable indicating whether the tweets were published before (coded 0) or after the intervention (coded 1). Therefore, β_2 indicates the intercept shift following the character limit intervention on the feature value, while β_3 shows the slope change after the character limit intervention. Including quadratic or cubic time trends (described in the Supplementary Materials) did not substantively change our results.

Although we used the standard ITS model, we focused on β_2, i.e., the intercept shift, and not the slope shift (β_3). The slope, in part, indicated whether any post-treatment intercept shift returns to baseline after some time. This reversion to the mean may even be

unrelated to the character limit change. The OLS estimates for the slope change are reported in Table 6 of the Supplementary Materials.

Using this regression model, we examined the effect of character limit intervention on various uncivil and deliberative attributes $Y_{feature}$. The regression results in tables show the average of the 100 iterations, while the LOESS figures show one random iteration for purposes of illustration. We focused on the 100 days before and 100 days after the character switch (i.e., bandwidth set to 100). For expository purposes, we also present a simple difference in means in the feature prominence pre and post- the intervention.

Additional specifications

Effects among Compliers. Using the ITS framework, we also compared the pre-intervention levels among the entire sample to post-intervention levels only among tweets that contained more than 145 characters (compliers) and among those that contained less than 145 characters (non-compliers). We expected that the intervention would have a significant effect on the uncivil and deliberative characteristics of the tweets by the compliers.

Instrumental Variable (IV) estimates. Since not every person "complies" with treatment (uses more than 145 characters), we also used a fuzzy regression discontinuity framework which considered the intervention day as a discontinuity and time as the running variable (Hausman & Rapson, 2018). The intervention exogenously increases the probability that someone uses more than 145 characters. Our causal estimand in this model is the Local Average Treatment Effect, i.e., the effect among those who used more than 145 characters. A dummy indicating whether the tweet was posted before (coded 0) or after (coded 1) November 7 was the instrumental variable. The number of characters in a tweet was the endogenous variable (coded 0 if less than 145 characters; 1 if more than 145 characters).

Within Subject effects. Since some people who are more likely to tweet different types of messages may be more likely to use more than 145 characters, we also tested our

results for any evidence of self-selection bias. We limited the dataset to those subjects that tweeted before and after the intervention and included subject fixed effects, which effectively examines the effect of the character limit change within subjects, washing away time-invariant effects.

Bandwidth effects. To test whether our results were an artifact of the bandwidth chosen in our ITS models, we also replicated our ITS models with different bandwidths of data.

Placebo effects. We replicated our analysis on a dataset of tweets posted in 2016, using the same data collection method as above to first identify 66,927 replies to the same U.S. politicians in the 100-day bandwidth before and after November 7, 2016. We replicated our experimental analysis to calculate the difference in means and a linear model specification. We report the average standardized coefficients and standard errors for 100 iterations, conducted on an average of 400 daily observations sampled with replacement from the dataset.

Results

As a manipulation check, we first assessed whether intervention affected the number of words used per tweet. The average number of words per tweet increased significantly after the character limit intervention (β: .71, p = $3.4e^{-5}$). This effect did not decline over time (coefficient of interaction: .003, p = .17).

The core analysis is performed on 358,242 tweets, of which 146,878 were posted after the intervention (Twitter's character limit change). 99.09% of all the tweets collected were found to be relevant to politics, which gives us confidence that our findings apply to describe online political discussions. The distribution of different uncivil and deliberative attributes and their intercorrelation is provided in Table 1 and Figure 1 of the Supplementary Materials. Compliers posted 70,440 tweets (48.0%) after the intervention. Figure 1 shows the change in the uncivil and deliberative attributes in the period leading up to and after the intervention, where the solid curves were generated by locally weighted

regression of the attributes on sequential day numbers, with no adjustment for covariates. Post-intervention, the blue curve reflects the trend for compliers while the red curve depicts the non-compliers.

This visualization provides the first set of clues regarding the effect of the intervention on online political discourse. Post-intervention, compliers (blue) and non-compliers (red) appear to be markedly different across most of the uncivil and deliberative characteristics. No discontinuities at the intervention are observed for offensiveness. Incivility, swear words and informal words show a discontinuous decrease on the day of the intervention. Politeness, justification, constructiveness, and reciprocity show a discontinuous increase, and empathy and respect decreased discontinuously.[3] In considering the slopes of these curves (the β_3 coefficients, reported in Table 6 of the Supplementary Materials), the findings suggest that politeness shows a temporary level change and it would revert to the pre-intervention mean over time. On the other hand, the decrease in the use of uncivil words, and empathy and respect reflect both a level and a slope change, and their presence in tweets is likely to further decrease over time.

We formally test these difference in Table 3. These OLS estimates are fit to the daily trends for a bandwidth of 100 days and feature two functional forms. Aggregating by day, Column 1 presents an estimate of the mean difference in the attributes before and after the intervention. Column 2 fits a linear function to the data on either side of the intervention date. Columns 3 and 4 provide the results considering all the tweets pre-intervention and separating the effects among compliers and non-compliers post-intervention. Column 5 provides the local average treatment effects among compliers following a fuzzy regression discontinuity design implementation. Columns 6 and 7 report the estimates after considering subject fixed effects, for those in our dataset who tweeted both pre- and post-intervention.[4]

To answer our research question, we observe that two out of four measures of uncivil

behavior significantly changed post-intervention across all specifications, to make the discussion more civil overall. Offensiveness only differed in the first (and least credible) specification. Average treatment effects reporting reduced uncivil behavior among compliers are significant in three of the four specifications. Within-subject effects among compliers reporting reduced uncivil behavior are significant in two of four specifications: lower offensiveness (Column 3: β = -.11, p = 3.86e^{-8}; Column 5: β = -.12, p= 2.73e^{-12}) and lower proportions of uncivil words (Column 3: β = -.27, p = 7.6e^{-32}; Column 5: β = -.23, p = 4.05e^{-54}; Column 7: β = -.21, p = 1.20e^{-15}). While the overall use of swear words did not significantly change, compliers were less likely to use them after the intervention (Column 3: β = -.07, p = 1.47e^{-5}; Column 5: β = -.10, p = 6.92e^{-10}).

To test H1, consider the results for informality. Compliers were less likely to use informal language after the character-limit change (Column 3: β = -.12, p = 1.49e^{-8}; Column 5: β = -.13, p = 3.08e^{-15}; Column 7: β = -.12, p = 5.98e^{-11}). Politeness significantly increased among compliers (Column 3: β = .08, p = 4.0e^{-4}; Column 5: β = .08, p = 3.86e^{-11}) and was marginally significant at the within-subject level (Column 7: β = .04, p = .07). While politeness also increased among non-compliers, the effect was 33% larger among the compliers than non-compliers. In summary, our findings show that with the increase in the character limit per post, political discussions on Twitter became more civil. Therefore, *H1 is supported by the results.*

To test H2a, consider the results for justification and constructiveness. Across all model specifications, compliers were more likely to use justifications after the character limit change (Column 3: β = .08, p = 1.11e^{-4}; Column 5: β = .06, p=3.36e^{-4}) which was marginally significant at the within subject level (Column 7: β = .04, p=.08). They were also more likely to write constructive tweets across all the specifications (Column 3: β = .16, p = 1.86e^{-9}; Column 5: β = .13, p = 2.32e^{-16}; Column 7: β = .14, p = 1.5e^{-6}) in political discussions after the intervention.

To text H2b, consider the results for reciprocity, and empathy and respect. Across all

model specifications, compliers were more likely to manifest reciprocity in their political discussions after the character limit change (Column 3: β = .18, p = 6.08e^{-11}; Column 5: β = .19, p = 4.209e^{-28}; Column 7: β = .19, p = 1.12e^{-16}). However, they were also *less* likely to show empathy and respect (Column 3: β = -.08, p = 5.2e^{-4}; Column 5: β = -.04, p= .02; Column 7: β = -.07, p = 6.7e^{-4}). Thus, while interpersonal engagement increased, there was a decrease in the general empathy and respect towards others.

Taking together the results from H2a and H2b, we can say that *the results partially support H2.* In terms of the analytical processes of political discussions, there is a promising increase in the use of facts and data to support arguments. At the same time, however, the quality of socializing appears to have decreased, with writing becoming more formal yet less sensitive to other viewpoints.

Interpretation. We note then that the post-intervention differences for political content are small at the tweet level and the subject level (Columns 5 and 7 of Table 3). To interpret the effect of the character limit change on the political discussion, we report the effects when the data is mean-centered at the day level and found that in many cases, the day-level shift among compliers is in the range of .7 - 1.3%.[5] Consequently, we expect that across millions of tweets and people, small within-subject effects can lead to a large difference in the overall characteristics of the political conversation, as visualized in the discontinuities in Figure 1.

Robustness and Placebo Tests

Figure 2 displays the estimated treatment effects using the linear functional form with various bandwidths N=[1, 100]. The left panel shows the result by fitting the linear model to all observations, while the right panel considers only compliers post-treatment. Shaded points show significant estimates at three different levels (p <= .05, p <= .01, or p <= .001). These trends show that the choice of bandwidth affects the significance of effects when there is a smaller sample size. Importantly, our findings do not appear to be contingent on the choice of bandwidth because in most of the cases the direction of the

effect remains consistent (i.e., either all effects are negative, or all effects are positive) across large and small bandwidths.

For the placebo dataset collected in November 2016, when there was no real intervention or increase in the Twitter character limit, we observed that the results do not repeat the main trends of Table 3 post- the placebo treatment. These results are reported in Table 7 of the Supplementary Materials.

Discussion and Recommendations

While many past studies have examined the characteristics of informal, online political conversations, these studies have often been limited to cross-sectional observational settings or artificial lab settings. The current study takes advantage of the Twitter character limit change and employs a quasi-natural experiment approach, thus strengthening previous scholars' arguments specifically about the deliberative qualities of everyday political discussions, and more generally about the importance of affordances in the study of political communication. No previous study has previously coded the deliberative characteristics of a dataset of this size, either manually or automatically; thus this study offers the first benchmark of the deliberative characteristics of everyday political discussions.

Our findings offer insights into how affordances can manipulate political deliberation and interactivity. They are robust across all specifications, and we expect them to generalize to discussions on other social media platforms as well. The statistics for incivility are distributionally similar to those reported in other corpora of political comments by Theocharis et al. (2016), Oz et al. (2018) and Muddiman et al. (2018). While we expect there to be individual variances according to the issues people are discussing, or to whom the replies were addressed, we do expect that our findings of the *change* in uncivil and deliberative characteristics do generalize across all issues. It would be interesting to consider how a change in other affordances, e.g., anonymity, visibility of social cues, placement of action buttonsMatias (2019) could change online social norms, or facilitate

conversations across political lines of difference (Settle & Carlson, 2019). In future work, we plan to explore the interplay of these and other affordances for their impact on online political discussions.

As with all quasi-experimental designs, we cannot discount all threats to internal validity (Shadish, Cook, & Campbell, 2002). In particular, what Shadish et al. (2002) call internal validity threat due to history remains possible. November 8, 2017, is the first anniversary of the 2016 United States presidential election, which is very close to the date of Twitter character limit change. The event might have briefly changed the uncivil and deliberative characteristics of the tweets around that time. However, the effect would eventually fade away, while the new 280 character limit would not. Moreover, as our analysis shows, the effect of the character limit change has often remained months after the event. Furthermore, a placebo test conducted using the data of November 2016 reaffirms the validity of our conclusions.

The findings of our study have practical implications for future work. Firstly, our study highlights the need to conduct more time series analyses with social media data to better understand evolving social norms on social media platforms, and their impact on the public sphere. Given the nature of our sample, our findings explain the change in individual tweets; however, inferences at the conversation level would require all the tweets and replies in a single thread. Although we employed the Twitter streaming API in our study, which only returns a 1% random sample of all public tweets, we recommend scholars to use Twitter's firehose stream when possible, which charges based on the date range and the number of tweets requested by researchers. Studies have debated on the sampling validity of the streaming API for quite some time (see González-Bailón, Wang, Rivero, Borge-Holthoefer, & Moreno, 2014; Morstatter, Pfeffer, & Liu, 2014; Morstatter, Pfeffer, Liu, & Carley, 2013, for details). However, a time-series analysis requires a wide band of data, which often makes it monetarily unrealistic to use the full-access, yet expensive, alternative.

Secondly, while this study focuses on the political discussions in the United States (specifically, replies addressed to US politicians), it can be expected that Twitter's affordances also influence conversations on other topics, and in other social and cultural contexts. We recommend that the limitations of the sampling issue and the historic threat should be tested by extending this research project to other contexts. We invite colleagues to replicate our study in other countries and languages to examine whether the effect of affordance change on political discussions is contingent on cultural and political factors. Further research can also be conducted in experimental settings to mitigate environmental confounders and further our understanding of the dynamics between technological affordances and human communication.

Conclusion

Through this study, we have shown how the design decisions made by computer engineers and technology companies might have a profound impact on democratic processes (Forestal, 2017). In light of recent events where communication technology had a detrimental effect on democracy, such as the compromise of personal data around the 2016 US presidential election (Cadwalladr & Graham-Harrison, 2018), rampant misinformation on social media (Del Vicario et al., 2016), and some evidence of online echo chambers (at least on social media) (e.g., Barberá, Jost, Nagler, Tucker, & Bonneau, 2015; Flaxman, Goel, & Rao, 2016; Sunstein, 2017), we contend that technology companies have a responsibility to make communication platforms more friendly for political discussions. However, this too should be implemented with careful consideration of users' rights; this study does not advocate technological determinism as a catch-all solution for improving the online political sphere.

The increasing accessibility of massive datasets and the development of new tools for computational and linguistic analyses have enabled scholars to use more sophisticated techniques to model computer-mediated communication; this also has broader implications for the study of political communication. Methodological frameworks to identify causal

effects, such as instrumental variables and regression discontinuity designs, are also becoming increasingly popular in social science disciplines (e.g., Dunning, 2008). We hope that this study encourages others in communication to combine computational approaches with the tools of causal inference to identify media effects in the digital world.

Many have argued that popular social media platforms might not be suitable for constructive discussions, citing the negative human behavior demonstrated online such as toxicity, incivility, and lack of empathy (e.g., Anderson, Brossard, Scheufele, Xenos, & Ladwig, 2014; Baek, Wojcieszak, & Delli Carpini, 2012; Coe, Kenski, & Rains, 2014). Some have suggested that a solution to online incivility lies in exposing people to different content or exhortations to act civil (e.g., Kim, 2015; Munger, 2017). This study corroborates the perspective that deconstructing social media platforms to identify what exactly following a 'mixture of attributes' approach, o the attributes of social media platforms technological affordances is also a way forward for studying online political communication (Eveland, 2003; Walther et al., 2005). We present a quasi-experiment on the impact of a technological affordance change on political discussion. While Twitter's implementation of a character limit change to double the length of tweets led to less uncivil political discussions and more deliberative political discussions, it also decreased empathy and respect among the discussants.

The findings assuage some doubts about the quality of online political discussions or its potential as a public sphere. Firstly, over 99% of all messages were labeled as relevant to politics, which validates our and Theocharis et al.'s (2016) sampling method, and also belies the general perception of these discussions as mere chatter. Secondly, only about 3% of the messages are labeled uncivil, which again contradicts the general perception of online political discussions as only toxic and frustrating.

Our study highlights the potential impact of affordances on online political communication in general. Firstly, it hints at how affordances can trigger a change in media effects, because we observe an increase in information sharing behavior (justification

and constructiveness; precisely, the sharing of facts) after the change of affordances. Another implication is the impact of affordances on the perceived credibility of online messages. Findings support the expectation that with the increased character limit, users have more space to clarify their meaning and provide arguments or evidence in their support(Chen, 2017; Theocharis et al., 2016), and possibly be perceived as more credible and authoritative than before.

Secondly, the findings suggest that a change in affordances can also trigger a change in online social relationships. While the character limit improved overall civility in political discussion, the decrease in empathy and respect among compliers is a cause for concern. That is to say, while people were more polite after the character limit change, they were less likely to be empathetic or respectful to other people's comments. The findings offer a different dimension to measure the tone of deliberation. Previous studies have implied that less respectful messages may be considered more persuasive (Fridkin & Kenney, 2008) and more authentic (Benson, 2011), perhaps signaling sincerity and the speaker's high stakes in the disagreement (Benson, 2011). Future work could explore whether changing interactivity affordances could help to include more voices through friendlier responses, or change the perceptions of authenticity and trust in online political discussions.

Finally, there is cause for concern that a change in affordances creates new social norms and structural inequalities in the online public sphere. Our findings show that the effects of Twitter's new affordance are largely limited to those who self-select into applying them in their daily communication. Therefore, the overall improvement in the argumentative quality of political discussions on Twitter may only matter for those social media users who are already politically engaged. Scholars have shown that elite users—already more engaged, articulate and authoritative—lead the political rhetoric in the *dominant*, rather than the public sphere of debate (e.g., Fraser, 1990; McGregor, 2018; Papacharissi, 2004).

While this study offers a starting point, much needs to be done to examine whether

these findings would lead to changes in overall political engagement, or generalize to other political or language norms. Political engagement on social media is especially important in countries with a greater reliance on internet election campaigning (Ahmed & Skoric, 2014), and in countries with restricted press freedom or a more authoritarian government (Ahmed & Cho, 2019). In terms of generalizability to other cultures and language norms, we should note that as per the original conceptualization, an affordance denotes the *relation* between the environment and the observer (**?**). Twitter noted that only 0.4% of Japanese tweets hit the 140-character limit whereas 9% of English tweets do. In cross- or multi-lingual contexts, there is often more than a character limit at play, and even a universal character limit actually has rather different effects in different languages (Liao, Fu, & Hale, 2015). We thus encourage other researchers to extend our research and examine the impact of affordances on online social norms in other environments and languages.

To summarize, while an online political utopia may remain a pipe-dream, this study has shown that we can make the current environment more hospitable to democratic discourse. While platforms are understandably cautious when it comes to censoring or promoting certain content or users, this study indicates that they have other tools that will promote healthy political discourse. Internet platforms are always A/B testing their products to increase profits and revenue, and to improve user engagement and satisfaction (Hindman, 2018). They may also want to test the impact of such design changes on discussion health. Of course, platforms will need to balance their profit motives for their public interest motives when making these decisions, but we hope that the latter will play a significant role in this calculus.

Footnotes

[1]Code and Data deposition: The code and data reported in this paper have been deposited in the Open Science Framework, https://osf.io/u2nfp/

[2]See Tausczik and Pennebaker (2010) for more details about validation, and Pennebaker, Boyd, et al. (2015) for details on LIWC's construction and inter-coder reliability.

[3]Incivility solely measures the presence of abuses, extremist language or vulgarity. Justification solely comprises evidence of facts, data or links in the tweets, and Constructiveness solely comprises the 'fact-checking' aspect. Although in the annotation task, tweets were labeled for the presence or absence of the individual facets of inciviity, justification and cosntructiveness respectively, there were not enough positive instances of other aspects to warrant training a classifier.

[4] A post hoc analysis revealed that for a margin of error of 4% in our within-subject analysis, and given the number of subjects (N=7382) and the smallest effect size (0.05) in the present study, we obtain a statistical power of 0.99.

[5]The results are reported in Table 6 of the Supplementary Materials.

References

Ahmed, S., & Cho, J. (2019). The internet and political (in) equality in the arab world: A multi-country study of the relationship between internet news use, press freedom, and protest participation. *New Media & Society*, *21*(5), 1065–1084.

Ahmed, S., & Skoric, M. M. (2014). My name is khan: the use of twitter in the campaign for 2013 pakistan general election. In *2014 47th hawaii international conference on system sciences* (pp. 2242–2251).

Almeida, T. G., Souza, B. , Nakamura, F. G., & Nakamura, E. F. (2017). Detecting hate, offensive, and regular speech in short comments. In *Proceedings of the 23rd Brazillian Symposium on Multimedia and the Web - WebMedia '17* (pp. 225–228). Gramado, Brazil: ACM Press. doi: 10.1145/3126858.3131576

Althoff, T., Danescu-Niculescu-Mizil, C., & Jurafsky, D. (2014). How to ask for a favor: A case study on the success of altruistic requests. In *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media* (pp. 12–21).

Anderson, A. A., Brossard, D., Scheufele, D. A., Xenos, M. A., & Ladwig, P. (2014). The "nasty effect:" Online incivility and risk perceptions of emerging technologies. *Journal of Computer-Mediated Communication*, *19*(3), 373–387. doi: 10.1111/jcc4.12009

Ausserhofer, J., & Maireder, A. (2013). National politics on Twitter: Structures and topics of a networked public sphere. *Information, Communication & Society*, *16*(3), 291–314. doi: 10.1080/1369118X.2012.756050

Baek, Y. M., Wojcieszak, M., & Delli Carpini, M. X. (2012). Online versus face-to-face deliberation: Who? Why? What? With what effects? *New Media & Society*, *14*(3), 363–383. doi: 10.1177/1461444811413191

Bail, C. A., Argyle, L. P., Brown, T. W., Bumpus, J. P., Chen, H., Hunzaker, M. B. F., ... Volfovsky, A. (2018). Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, *115*(37), 9216–9221. doi: 10.1073/pnas.1804840115

Barber, B. R. (1984). *Strong democracy: Participatory politics for a new age*. Berkeley, CA: University of California Press.

Barberá, P., Jost, J. T., Nagler, J., Tucker, J. A., & Bonneau, R. (2015). Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological Science*, *26*(10), 1531–1542. doi: 10.1177/0956797615594620

Benson, T. W. (2011). The rhetoric of civility: Power, authenticity, and democracy. *Journal of Contemporary Rhetoric*, *1*(1), 22–30.

Berry, J. M., & Sobieraj, S. (2013). *The outrage industry: Political opinion media and the new incivility*. Oxford, UK: Oxford University Press.

Cadwalladr, C., & Graham-Harrison, E. (2018). *Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach.* Retrieved from `https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election`

Chen, G. M. (2017). *Online incivility and public debate: Nasty talk*. Cham, Switzerland: Palgrave Macmillan.

Coe, K., Kenski, K., & Rains, S. A. (2014). Online and uncivil? Patterns and determinants of incivility in newspaper website comments. *Journal of Communication*, *64*(4), 658–679. doi: 10.1111/jcom.12104

Cook, F. L., Carpini, M. X. D., & Jacobs, L. R. (2007). Who deliberates? Discursive participation in America. In S. W. Rosenberg (Ed.), *Deliberation, participation and democracy* (pp. 25–44). London, UK: Palgrave Macmillan. doi: 10.1057/9780230591080_2

Danescu-Niculescu-Mizil, C., Sudhof, M., Jurafsky, D., Leskovec, J., & Potts, C. (2013). A computational approach to politeness with application to social factors. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 250–259). Sofia, Bulgaria: Association for Computational Linguistics.

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In *Proceedings of the Eleventh International AAAI Conference on Web and Social Media* (pp. 512–515).

Davis, A. (2010). New media and fat democracy: The paradox of online participation. *New Media & Society*, *12*(5), 745–761. doi: 10.1177/1461444809341435

Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., ... Quattrociocchi, W. (2016). The spreading of misinformation online. *Proceedings of the National Academy of Sciences*, *113*(3), 554–559. doi: 10.1073/pnas.1517441113

Dunning, T. (2008). Improving causal inference: Strengths and limitations of natural experiments. *Political Research Quarterly*, *61*(2), 282–293. doi: 10.1177/1065912907306470

Effing, R., van Hillegersberg, J., & Huibers, T. (2011). Social media and political participation: Are Facebook, Twitter and YouTube democratizing our political systems? In E. Tambouris, A. Macintosh, & H. de Bruijn (Eds.), *Electronic participation* (pp. 25–35). Berlin, Heidelberg: Springer.

Esteve Del Valle, M., Sijtsma, R., & Stegeman, H. (2018). Social media and the public sphere in the Dutch parliamentary Twitter network: A space for political deliberation? Hamburg, Germany: ECPR General Conference.

Evans, S. K., Pearce, K. E., Vitak, J., & Treem, J. W. (2017). Explicating affordances: A conceptual framework for understanding affordances in communication research. *Journal of Computer-Mediated Communication*, *22*(1), 35–52. doi: 10.1111/jcc4.12180

Eveland, W. P. (2003). A "mix of attributes" approach to the study of media effects and new communication technologies. *Journal of Communication*, *53*(3), 395–410. doi: 10.1111/j.1460-2466.2003.tb02598.x

Eveland, W. P., Morey, A. C., & Hutchens, M. J. (2011). Beyond deliberation: New directions for the study of informal political conversation from a communication

perspective. *Journal of Communication*, *61*(6), 1082–1103. doi: 10.1111/j.1460-2466.2011.01598.x

Faraj, S., & Azad, B. (2012). The materiality of technology: An affordance perspective. In P. M. Leonardi, B. A. Nardi, & J. Kallinikos (Eds.), *Materiality and organizing: Social interaction in a technological world* (pp. 237–258). Oxford, UK: Oxford University Press.

Flaxman, S., Goel, S., & Rao, J. M. (2016). Filter bubbles, echo chambers, and online news consumption. *Public Opinion Quarterly*, *80*(S1), 298–320. doi: 10.1093/poq/nfw006

Forestal, J. (2017). The architecture of political spaces: Trolls, digital media, and Deweyan democracy. *American Political Science Review*, *111*(1), 149–161. doi: 10.1017/S0003055416000666

Fraser, B. (1990). Perspectives on politeness. *Journal of Pragmatics*, *14*(2), 219–236. doi: 10.1016/0378-2166(90)90081-N

Freelon, D. G. (2010). Analyzing online political discussion using three models of democratic communication. *New Media & Society*, *12*(7), 1172–1190. doi: 10.1177/1461444809357927

Fridkin, K. L., & Kenney, P. J. (2008). The dimensions of negative messages. *American Politics Research*, *36*(5), 694–723. doi: 10.1177/1532673X08316448

Friess, D., & Eilders, C. (2015). A systematic review of online deliberation research. *Policy & Internet*, *7*(3), 319–339. doi: 10.1002/poi3.95

Gastil, J. (2008). *Political communication and deliberation*. Los Angeles, CA: SAGE Publications.

Gibson, J. J. (1977). The theory of affordances. In R. Shaw & J. Bransford (Eds.), *Perceiving, acting, and knowing: Toward an ecological psychology* (pp. 67–82). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Gligorić, K., Anderson, A., & West, R. (2018). How constraints affect content: The case of Twitter's switch from 140 to 280 characters. In *Proceedings of the Twelfth*

International AAAI Conference on Web and Social Media (pp. 596–599).

González-Bailón, S., Wang, N., Rivero, A., Borge-Holthoefer, J., & Moreno, Y. (2014). Assessing the bias in samples of large online networks. *Social Networks*, *38*, 16–27. doi: 10.1016/j.socnet.2014.01.004

Grant, W. J., Moon, B., & Busby Grant, J. (2010). Digital dialogue? Australian politicians' use of the social network tool Twitter. *Australian Journal of Political Science*, *45*(4), 579–604. doi: 10.1080/10361146.2010.517176

Greeno, J. G. (1994). Gibson's affordances. *Psychological Review*, *101*(2), 336–342.

Groshek, J., & Cutino, C. (2016). Meaner on mobile: Incivility and impoliteness in communicating contentious politics on sociotechnical networks. *Social Media + Society*, *2*(4), 1–10. doi: 10.1177/2056305116677137

Habermas, J. (1984). *The theory of communicative action* (Vol. 2). Boston, MA: Beacon Press.

Habermas, J. (1990). *Moral consciousness and communicative action*. Cambridge, MA: MIT Press.

Habermas, J., Lennox, S., & Lennox, F. (1974). The public sphere: An encyclopedia article (1964). *New German Critique*(3), 49–55. doi: 10.2307/487737

Halpern, D., & Gibbs, J. (2013). Social media as a catalyst for online deliberation? Exploring the affordances of Facebook and YouTube for political expression. *Computers in Human Behavior*, *29*(3), 1159–1168. doi: 10.1016/j.chb.2012.10.008

Hausman, C., & Rapson, D. S. (2018). Regression discontinuity in time: Considerations for empirical applications. *Annual Review of Resource Economics*, *10*(1), 533–552. doi: 10.1146/annurev-resource-121517-033306

Himmelroos, S. (2017). Discourse quality in deliberative citizen forums – A comparison of four deliberative mini-publics. *Journal of Public Deliberation*, *13*(1), Article 3.

Hindman, M. S. (2018). *The Internet trap: How the digital economy builds monopolies and undermines democracy*. Princeton, NJ: Princeton University Press.

Hopmann, D. N., Matthes, J., & Nir, L. (2015). Informal political conversation across time and space: Setting the research agenda. *International Journal of Public Opinion Research*, *27*(4), 448–460. doi: 10.1093/ijpor/edv043

Jaidka, K., Guntuku, S. C., Buffone, A., Schwartz, H. A., & Ungar, L. H. (2018). Facebook vs. Twitter: Cross-platform differences in self-disclosure and trait prediction. In *Proceedings of the Twelfth International AAAI Conference on Web and Social Media* (pp. 141–150).

Janssen, D., & Kies, R. (2005). Online forums and deliberative democracy. *Acta Politica*, *40*(3), 317–335. doi: 10.1057/palgrave.ap.5500115

Jongeling, R., Sarkar, P., Datta, S., & Serebrenik, A. (2017). On negative results when using sentiment analysis tools for software engineering research. *Empirical Software Engineering*, *22*(5), 2543–2584. doi: 10.1007/s10664-016-9493-x

Joyce, C. K. (2009). *The blank page: Effects of constraint on creativity* (Doctoral Dissertation, University of California, Berkeley, Berkeley, CA). Retrieved from `http://dx.doi.org/10.2139/ssrn.1552835`

Kim, Y. (2015). Does disagreement mitigate polarization? How selective exposure and disagreement affect political polarization. *Journalism & Mass Communication Quarterly*, *92*(4), 915–937. doi: 10.1177/1077699015596328

Lee, E.-J., & Oh, S. Y. (2012). To personalize or depersonalize? When and how politicians' personalized tweets affect the public's reactions. *Journal of Communication*, *62*(6), 932–949. doi: 10.1111/j.1460-2466.2012.01681.x

Liao, H.-T., Fu, K.-w., & Hale, S. A. (2015). How much is said in a microblog?: A multilingual inquiry based on weibo and twitter. In *Proceedings of the acm web science conference* (p. 25).

Lin, H., & Qiu, L. (2013). Two sites, two voices: Linguistic differences between Facebook status updates and tweets. In D. Hutchison et al. (Eds.), *Cross-cultural design: Cultural differences in everyday life* (pp. 432–440). Berlin, Heidelberg: Springer.

Lindgren, S., & Lundström, R. (2011). Pirate culture and hacktivist mobilization: The cultural and social protocols of #WikiLeaks on Twitter. *New Media & Society*, *13*(6), 999–1018. doi: 10.1177/1461444811414833

Liu, N., & Zhang, X. (2013). The influence of group communication, government-citizen interaction, and perceived importance of new media on online political discussion. *Policy & Internet*, *5*(4), 444–461. doi: 10.1002/1944-2866.POI348

Lopez Bernal, J., Cummins, S., & Gasparrini, A. (2017). Interrupted time series regression for the evaluation of public health interventions: A tutorial. *International Journal of Epidemiology*, *46*(1), 348–355. doi: 10.1093/ije/dyw098

Matias, J. N. (2019). Preventing harassment and increasing group participation through social norms in 2,190 online science discussions. *Proceedings of the National Academy of Sciences*, *116*(20), 9785–9789.

McGregor, S. C. (2018). *Social (media) construction of public opinion by elites* (Doctoral Dissertation, University of Texas at Austin, Austin, TX). Retrieved from `http://hdl.handle.net/2152/67619`

Mendelberg, T. (2002). The deliberative citizen: Theory and evidence. *Political Decision Making, Deliberation and Participation*, *6*, 151–193.

Morstatter, F., Pfeffer, J., & Liu, H. (2014). When is it biased?: Assessing the representativeness of Twitter's Streaming API. In *Proceedings of the 23rd International Conference on World Wide Web - WWW '14 Companion* (pp. 555–556). doi: 10.1145/2567948.2576952

Morstatter, F., Pfeffer, J., Liu, H., & Carley, K. M. (2013). Is the sample good enough? Comparing data from Twitter's Streaming API with Twitter's Firehose. In *Proceedings of the Seventh International AAAI Conference on Web and Social Media* (pp. 400–408).

Muddiman, A., McGregor, S. C., & Stroud, N. J. (2018). (Re)claiming our expertise: Parsing large text corpora with manually validated and organic dictionaries. *Political*

Communication. doi: 10.1080/10584609.2018.1517843

Munger, K. (2017). Tweetment effects on the tweeted: Experimentally reducing racist harassment. *Political Behavior*, *39*(3), 629–649. doi: 10.1007/s11109-016-9373-5

Mutz, D. C., & Reeves, B. (2005). The new videomalaise: Effects of televised incivility on political trust. *American Political Science Review*, *99*(1), 1–15. doi: 10.1017/S0003055405051452

Neblo, M. A., Esterling, K. M., Kennedy, R. P., Lazer, D. M., & Sokhey, A. E. (2010). Who wants to deliberate—And why? *American Political Science Review*, *104*(3), 566–583. doi: 10.1017/S0003055410000298

Nithyanand, R., Schaffner, B., & Gill, P. (2017). Online political discourse in the Trump era. *arXiv:1711.05303*.

Olteanu, A., Talamadupula, K., & Varshney, K. R. (2017). The limits of abstract evaluation metrics: The case of hate speech detection. In *Proceedings of the 2017 ACM on Web Science Conference - WebSci '17* (pp. 405–406). Troy, NY: ACM Press. doi: 10.1145/3091478.3098871

Oz, M., Zheng, P., & Chen, G. M. (2018). Twitter versus Facebook: Comparing incivility, impoliteness, and deliberative attributes. *New Media & Society*, *20*(9), 3400–3419. doi: 10.1177/1461444817749516

Papacharissi, Z. (2004). Democracy online: Civility, politeness, and the democratic potential of online political discussion groups. *New Media & Society*, *6*(2), 259–283. doi: 10.1177/1461444804041444

Papacharissi, Z. (2010). *A private sphere: Democracy in a digital age*. Malden, MA: Polity.

Pennebaker, J. W., Booth, R. J., Boyd, R. L., & Francis, M. E. (2015). *Linguistic Inquiry and Word Count: LIWC 2015.* Austin, TX: Pennebaker Conglomerates.

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of LIWC2015* (Tech. Rep.). Austin, TX: University of

Texas at Austin.

Rowe, I. (2015). Civility 2.0: A comparative analysis of incivility in online political discussion. *Information, Communication & Society*, *18*(2), 121–138. doi: 10.1080/1369118X.2014.940365

Santana, A. D. (2014). Virtuous or vitriolic: The effect of anonymity on civility in online newspaper reader comment boards. *Journalism Practice*, *8*(1), 18–33. doi: 10.1080/17512786.2013.813194

Schkade, D., Sunstein, C. R., & Kahneman, D. (2000). Deliberating about dollars: The severity shift. *Columbia Law Review*, *100*, 1139–1176.

Settle, J. E., & Carlson, T. N. (2019). Opting out of political discussions. *Political Communication*, 1–21.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.

Shirky, C. (2008). *Here comes everybody: The power of organizing without organizations*. New York, NY: Penguin Press.

Steenbergen, M. R., Bächtiger, A., Spörndli, M., & Steiner, J. (2003). Measuring political deliberation: A discourse quality index. *Comparative European Politics*, *1*(1), 21–48. doi: 10.1057/palgrave.cep.6110002

Stromer-Galley, J. (2007). Measuring deliberation's content: A coding scheme. *Journal of Public Deliberation*, *3*(1), Article 12.

Stromer-Galley, J., & Martinson, A. M. (2009). Coherence in political computer-mediated communication: Analyzing topic relevance and drift in chat. *Discourse & Communication*, *3*(2), 195–216. doi: 10.1177/1750481309102452

Stroud, N. J., Scacco, J. M., Muddiman, A., & Curry, A. L. (2015). Changing deliberative norms on news organizations' Facebook sites. *Journal of Computer-Mediated Communication*, *20*(2), 188–203. doi: 10.1111/jcc4.12104

Sundar, S. S. (2008). The MAIN model: A heuristic approach to understanding technology effects on credibility. In M. J. Metzger & A. J. Flanagin (Eds.), *Digital media, youth, and credibility* (pp. 73–100). Cambridge, MA: MIT Press.

Sunstein, C. R. (2017). *#Republic: Divided democracy in the age of social media*. Princeton, NJ: Princeton University Press.

Tan, C., Niculae, V., Danescu-Niculescu-Mizil, C., & Lee, L. (2016). Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th International Conference on World Wide Web - WWW '16* (pp. 613–624). Montreal, Canada: ACM Press. doi: 10.1145/2872427.2883081

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, *29*(1), 24–54. doi: 10.1177/0261927X09351676

Theocharis, Y., Barberá, P., Fazekas, Z., Popa, S. A., & Parnet, O. (2016). A bad workman blames his tweets: The consequences of citizens' uncivil Twitter use when interacting with party candidates. *Journal of Communication*, *66*(6), 1007–1031. doi: 10.1111/jcom.12259

Theocharis, Y., Lowe, W., van Deth, J. W., & García-Albacete, G. (2015). Using Twitter to mobilize protest action: Online mobilization patterns and action repertoires in the Occupy Wall Street, Indignados, and Aganaktismenoi movements. *Information, Communication & Society*, *18*(2), 202–220. doi: 10.1080/1369118X.2014.948035

Towne, W. B., & Herbsleb, J. D. (2012). Design considerations for online deliberation systems. *Journal of Information Technology & Politics*, *9*(1), 97–115. doi: 10.1080/19331681.2011.637711

Tufekci, Z., & Wilson, C. (2012). Social media and the decision to participate in political protest: Observations from Tahrir Square. *Journal of Communication*, *62*(2), 363–379. doi: 10.1111/j.1460-2466.2012.01629.x

Walther, J. B., Gay, G., & Hancock, J. T. (2005). How do communication and technology

researchers study the Internet? *Journal of Communication*, *55*(3), 632–657. doi: 10.1111/j.1460-2466.2005.tb02688.x

Wojcieszak, M. E. (2010). 'Don't talk to me': Effects of ideologically homogeneous online groups and politically dissimilar offline ties on extremism. *New Media & Society*, *12*(4), 637–655. doi: 10.1177/1461444809342775

Wojcieszak, M. E., & Mutz, D. C. (2009). Online groups and political discourse: Do online discussion spaces facilitate exposure to political disagreement? *Journal of Communication*, *59*(1), 40–56. doi: 10.1111/j.1460-2466.2008.01403.x

Wyatt, R. O., Katz, E., & Kim, J. (2000). Bridging the spheres: Political and personal conversation in public and private spaces. *Journal of Communication*, *50*(1), 71–92. doi: 10.1111/j.1460-2466.2000.tb02834.x

Table 1

Dataset Description. Standard errors are in parantheses.

Number of	**Replies in the 1% Sample**	**After language filtering**	**After removing retweets**	**After removing early-access**
observations	2101856	1869481	398278	358242
Number of	**Unique users**	**Non-compliers**		**Compliers**
users	204902	62180		39246
Tweet	**Total number of words**	**Average words per tweet**		**Mean daily tweets**
characteristics	7668123	21.40 (12.6)	12.07	772.56 (479.76)

Table 2

The deliberative criteria used to measure the quality of political discussion.

Dimension	Operationalization	Source
Uncivil behavior		
Incivility	Whether a tweet contains abuses and vulgarity, threats, or exaggeration (1/0 binary coding)	This study
Offensiveness	Whether a tweet contains abuses and racial slurs (1/0 coding).	Davidson et. al, 2017
Uncivil words	The proportion of a tweet which comprises words from a curated list of offensive words (0 to 100 continuous values).	Muddiman, McGregor, & Stroud, 2018
Swear words	The proportion of a tweet which comprises words from a curated list of swear words (0 to 100 continuous values).	Linguistic Inquiry and Word Count (Pennebaker et. al, 2015)
Deliberative behavior		
Politeness	Likelihood that the tweet is polite (0 to 100 continuous values).	Stanford's Politeness classifier (Danescu-Niculescu-Mizil et. al, 2013)
Informal words	The proportionate use of informal language such as slang, fillers, swear words and netspeak (0 to 100 continuous values).	Linguistic Inquiry and Word Count (Pennebaker et. al, 2015)
Justification	Whether the tweet offers evidence in the form of feelings, experiences or facts (1/0 binary coding).	This study
Constructiveness	Whether the tweet offers a fact-check, search for common ground or a solution (1/0 binary coding).	This study
Reciprocity	Whether the tweet is intended to elicit an answer, information or feedback (1/0 binary coding).	This study
Empathy & Respect	Whether the tweet contains empathy or respect for others (1/0 binary coding).	This study

Table 3

OLS Estimates of the effect of platform change with different model and treatment specifications (bandwidth = 100 days)

	1 **Difference in Means**	2 **Linear**	3 Linear Compliers	4 **Linear Non-compliers**	5 **IV Estimates**	6 **Within All Subjects**	7 **Within Subject: Compliers**
Uncivil Behavior							
Incivility	.00^{+}	-.02	-.11***	.05*	-.12***	-.00^{+}	-.03
	(.009)	(.017)	(.015)	(.019)	(.017)	(.017)	(.017)
Offensiveness	.03***	.01	.014	.014	.02	.03	.05*
	(.009)	(.018)	(.019)	(.019)	(.016)	(.018)	(.019)
Uncivil words	-.11***	-.09***	-.27***	.04*	-.23***	-.10***	-.21***
	(.008)	(.016)	(.018)	(.018)	(.015)	(.024)	(.021)
Swear words	.02	-.02	-.07***	.02	-.10***	-.01	-.01
	(.009)	(.019)	(.016)	(.021)	(.017)	(.017)	(.017)
Deliberative Behavior							
Politeness	.04***	.07***	.08***	.05*	.11***	.04*	.04
	(.009)	(.019)	(.021)	(.019)	(.017)	(.020)	(.021)
Informal words	-.04***	-.04	-.12***	.03	-.13***	-.05***	-.12***
	(.009)	(.019)	(.018)	(.021)	(.017)	(.019)	(.018)
Justification	.01	.02	.08***	-.04***	.06***	.0^{+}	.04
	(.008)	(.017)	(.016)	(.020)	(.016)	(.020)	(.022)
Constructiveness	.07***	.07***	.16***	-.02	.13***	.05	.14***
	(.007)	(.016)	(.024)	(.021)	(.015)	(.024)	(.024)
Reciprocity	.08***	.08***	.18***	.0^{+}	.19***	.09***	.19***
	(.009)	(.019)	(.023)	(.017)	(.018)	(.022)	(.022)
Empathy & Respect	-.06***	-.02	-.08***	.03	-.04*	-.04	-.07***
	(.009)	(.018)	(.011)	(.018)	(.017)	(.022)	(.021)
Observations	200	200	200	200	180634	15271	7382

Note: *p<.05; ** p<.01; *** p<.001

Standard errors are shown in parentheses.

Figure 1. Daily trends of uncivil and deliberative behavior from January 2017 to March 2018 for political discussions on Twitter. The solid curve is generated by locally weighted (LOESS) regression of the uncivil and deliberative attributes on sequential day numbers, with no adjustment for covariates. Data were aggregated as day-level means here to facilitate visualization.

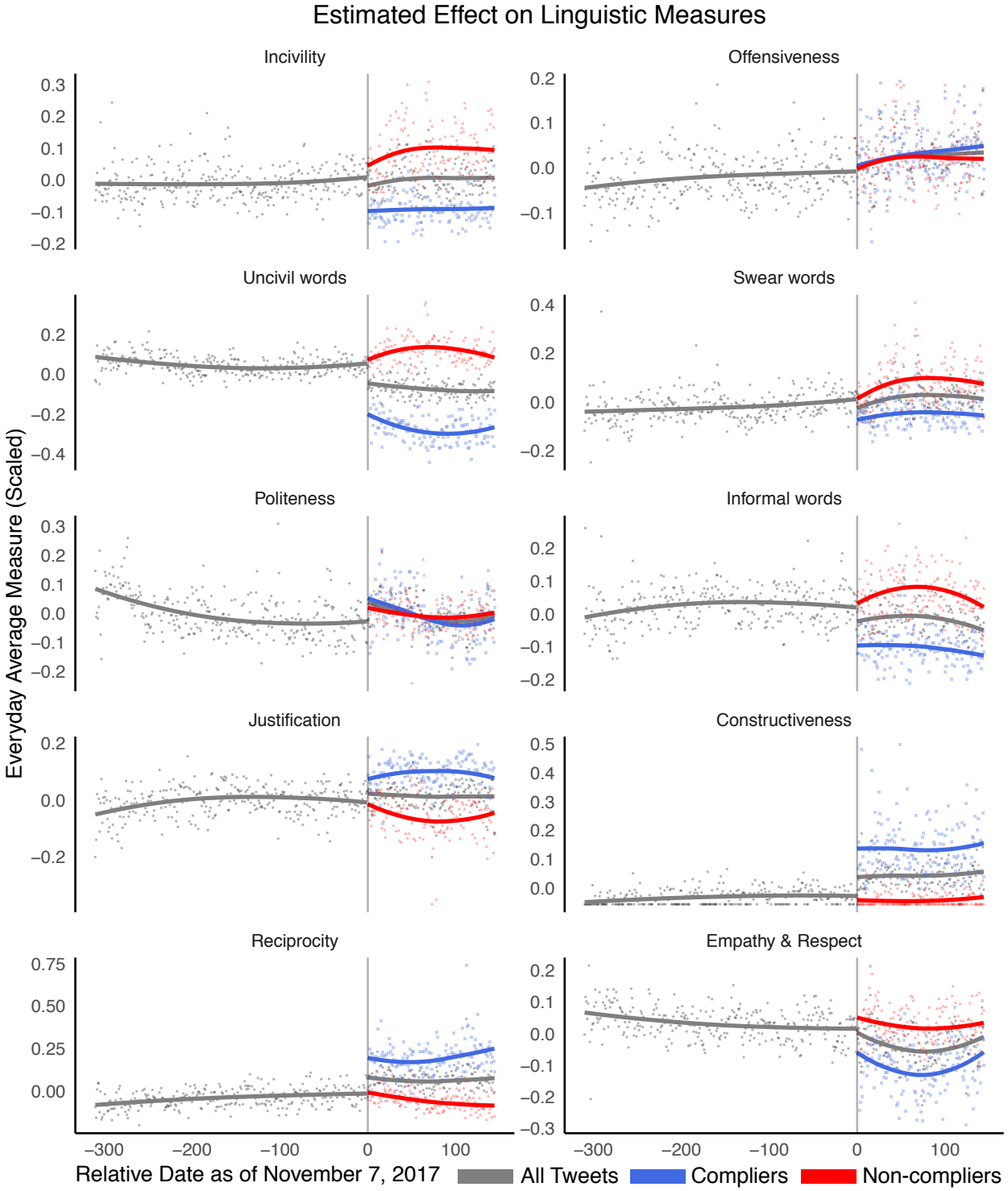

Figure 2. Estimates of the average change in uncivil and deliberative attributes after the intervention by fitting a linear model over (a) all observations (N = 358,242) and (b) compliers (N = 281,804), using various bandwidths [0,100] as plotted on the x-axis. Vertical lines denote 95% confidence intervals for each estimate.